DISSERTATION ABSTRACT

Voice Activity Detection Using Deep Neural Network

Graduate School of
Natural Science & Technology
Kanazawa University

Division of Electrical Engineering and Computer Science

Student ID Number: 1424042016
Name: Suci Dwijayanti
Chief Advisor: Masato MIYOSHI, Prof. Dr. Eng.
Date of Submission: December, 2017

Abstract

Voice activity detection (VAD) is used as a preprocessing for various speech applications to identify speech and non-speech periods in input signals. In this study, we propose a deep neural network (DNN)-based VAD method to detect such periods in noisy signals by utilizing dynamics that refer to the time-varying properties of speech signals. In the proposed method, the dynamics are highlighted by speech period candidates, which are calculated based on heuristic rules for trivial patterns of the first and second derivatives that characterize the starting and ending points of utterances. The candidates, together with log power spectra, are input into the DNN to determine speech periods. Experiments are conducted to evaluate the proposed method by using speech signals smeared with five types of noise (white, babble, factory, car and pink) with signal to noise ratios (SNRs) of 10, 5, 0 and -5 dB. The experimental results show that the proposed method is superior under all considered noise conditions. The proposed method improves the performance of DNN-based VAD using the log power spectra alone particularly in low SNRs or non-stationary noise. The addition of speech period candidates, which may highlight dynamics, provides positive information that contributes to find speech period correctly.

1 Introduction

Voice activity detection (VAD) is used as a preprocessing stage for various speech applications to identify speech and non-speech periods. In speech enhancement, for example in spectral subtraction, speech/non-speech detection is applied to identify the signal periods that contain only noise. This is useful for noise estimations, which are then used in the noise reduction process [1]. In digital cellular telecommunication systems, such as the universal mobile telecommunication systems (UMTS) [2], VAD is employed to detect non-speech frames and thus reduce average bit rates [3]. VAD may also improve the performance of speech recognition by identifying the boundaries of the speech to be recognized [4].

Because background noise is a challenging problem, selecting features that are discriminative for properties of speech and noise is an important aspect of the design of VAD algorithms [5]. In prior studies, simple acoustic features such as energy and zero crossing rates have been used to detect speech periods [6]. This type of technique is suitable for clean signals, and its performance is degraded under low signal-to-noise ratios (SNRs). Hence, various modifications of energy-based features, such as those described in [1] and [7], have been proposed to improve the VAD performance. Other acoustic features have also been examined and investigated to improve the VAD performance. For example, the VAD algorithm proposed in [8] measures the long-term spectral divergence between speech and noise. Periodic to aperiodic component ratios were employed in [9]. Pek *et al.* [10] used modulation indices of the modulation spectra of speech data. Kinnunen and Rajad [11] introduced likelihood ratio-based VAD method, in which speech and non-speech models are trained on an utterance-by-utterance basis using mel-frequency cepstral coefficients (MFCCs). Sohn *et al.* [12] proposed a method based on a Gaussian statistical model, in which a decision rule is derived from the mean of the likelihood ratios for individual frequency bands by assuming that the noise is known a priori. Davis *et al.* [13] proposed a scheme that incorporates a low-variance spectrum estimation technique and a method for determining an adaptive threshold based on noise statistics. These methods perform well under stationary noise; however, their performances are degraded under non-stationary noise. To improve the performance of VAD, machine learning methods have been explored. For instance, support vector machine (SVM) methods [14, 15, 16] and deep neural networks (DNNs) [17, 18, 19] have been found to be highly competitive with traditional VAD.

The great flexibility, deep and generative training properties of DNNs are useful in speech processing [20]. Espi *et al.* [21] utilized spectro-temporal features as the input to a convolutional neural network (CNN) to detect non-speech acoustic signals. Ryant *et al.* [22] utilized MFCCs as the input to a DNN to detect speech activity on YouTube. Mendelev *et al.* [23] proposed a DNN with a maxout activation function and dropout regularization to improve the VAD performance. Zhang *et al.* [24] attempted to optimize the capability of DNN-based VAD by combining multiple features, such as pitch, discrete Fourier transform (DFT), MFCCs, linear prediction coefficients (LPCs), relative-spectral perceptual linear predictive analysis (RASTA-PLP), and amplitude modulation spectrogram (AMS) along with their delta features, as the input to DNNs. However, choosing features as the input for a DNN is not a trivial problem. Research in automatic speech recognition has shown that raw features have the potential to be used as the input of a DNN, replacing "hand-crafted" features [25].

In this study, we first attempt to utilize raw features, i.e., log power spectra, to detect speech periods using a DNN. In our preliminary experiment, two findings are obtained. First, the performance of VAD using the log power spectra as the input of the DNN outperforms standard features, such as MFCCs and MFCCs combined with delta and delta-delta cepstra, for both clean and noisy speech signals. MFCCs lose some information from the speech signals; this may occur because of the use of discrete cosine transform (DCT) compression. Second, in the preliminary experiment, we find that the addition of delta and delta-delta cepstra to the MFCCs improves the VAD performance. Delta and delta-delta cepstra are features that express dynamics that refer to the time-varying properties of speech signals [26]. Thus, this result indicates that the dynamics may contribute to improving the VAD performance. Based on the second finding, we attempt to enhance the VAD performance based on the usage of log power spectra, adding the first and second derivatives of the log power spectra.

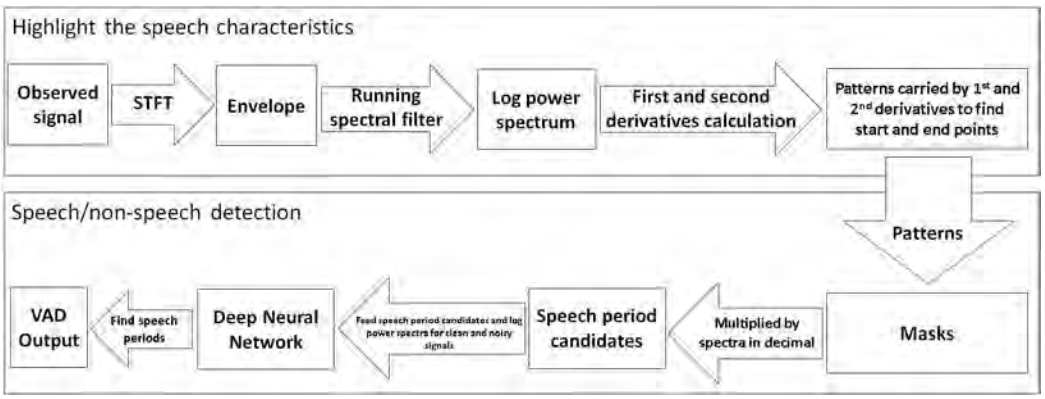


Figure 1: Outline of the proposed method

Figure 1 shows the outline of the proposed method. First, major speech characteristics are highlighted using a running spectral filter (RSF) [27]. Next, masks are composed using the first and second derivatives of the log power spectra of the

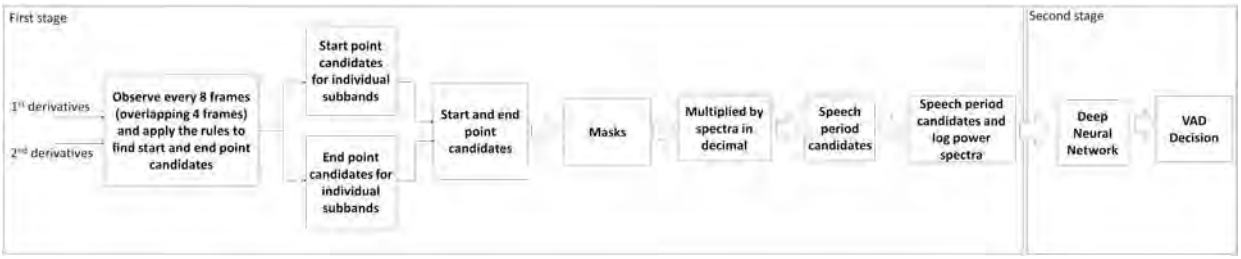


Figure 2: Block diagram of the proposed method

RSF output through heuristic rules. These masks, which consist of binary values, are then multiplied by spectra, expressed in decimal form, to obtain speech period candidates. Since not all subband signals might contribute to the VAD decision, we consider obtaining the speech period candidates for individual subbands. These speech period candidates, together with the log power spectra, are input into a DNN to obtain the VAD output. The experimental results show that the proposed method is superior to a DNN-based VAD method that utilizes the log power spectra alone.

The paper is organized as follows. In Section 2, we describe the proposed method for detecting speech periods using a combination of speech period candidates and log power spectra. In Section 3, the experimental results and a discussion of the results are provided. In Section 4, the conclusions of the paper are presented.

2 Proposed method

A speech signal can be analyzed by using a short-time Fourier transform (STFT) as follows:

$$X(m,k) = \sum_{n=\infty}^{\infty} x(n)h(m-n)W_k^{kn}, \tag{1}$$

where $x(n)$ is a speech signal, $h(n)$ is an analysis window, which is time reversed and shifted by m frames, k is a frequency bin variable, K is the number of frequency bins, and $W_K = \exp^{-j(\frac{2\pi}{K})}$. $X(m,k)$ can be further expressed as follows:

$$X(m,k) = |X(m,k)|e^{j\angle X(m,k)}. \tag{2}$$

As shown in several investigations [28, 29, 30, 31], the energy of clean speech signals is mostly concentrated within a modulation frequency range of 1 to 16 Hz. Hence, each subband envelope, $|X(m,k)|$, is filtered through an RSF to remove noise outside the modulation frequency range, and negative values in the filter output are replaced by zeros [32] .

A subband log power spectrum of the RSF output, $E(m,k)$, is expressed as

$$E(m,k) = 10\log_{10}(X_{rsf}(m,k))^2. \tag{3}$$

Hereafter, we call the log power spectrum of the RSF output as LPS-RSF. The first and second derivatives of the log power spectrum, $E(m,k)$, obtained through the above filtering, are calculated as follows:

$$\Delta_{-m}E(m) = E(m) - E(m-1) \tag{4}$$

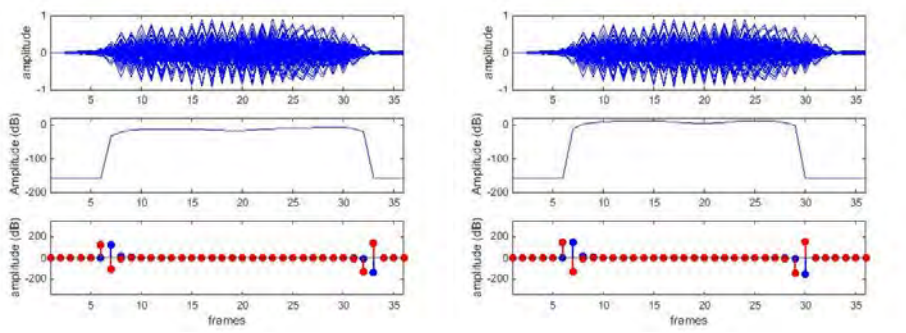
$$\begin{aligned}\Delta_m^2 E(m) &= \Delta_{+m}\Delta_{-m}E(m) \\ &= [E(m+1) - E(m)] - [E(m) - E(m-1)].\end{aligned} \tag{5}$$

These derivatives are used to obtain speech period candidates that highlight the dynamics in the LPS-RSF. These candidates are used together with the log power spectra, derived from Eq. (2), to detect speech periods in the DNN described in the next subsection. The detailed process is shown in Figure 2.

2.1 Speech period candidates

In spoken language, an utterance is a continuous piece of speech that has a start and an end and is separated from a successive utterance by a pause. Figure 3 shows the subband observations at 250 and 875 Hz of utterance /ha/ and the observations' first and second derivatives of the log power spectrum obtained using Eqs. (4) and (5). The frame size used to obtain this representation is 20 ms, which implies that each frame consists of 160 samples, and the analysis window is a Hamming window with a 10 ms frame shift.

As shown in Fig. 3, the starting and ending points of the utterance /ha/ may be identified from the patterns of the first and second derivatives. The starting and ending point candidates of utterance /ha/ in the subband at 250 Hz are located at frames 6 and 33, respectively. In contrast, in the subband at 875 Hz, the starting and ending point candidates are found to lie at frames 6 and 30, respectively. These observations indicate that not all subband signals may contribute to the VAD



(a) Subband at 250 Hz (b) Subband at 875 Hz

Figure 3: Subband observations of utterance /ha/, log power spectra, and their first and second derivatives. Blue and red lines represent first and second derivatives, respectively

(a) Starting point (b) Ending point

Figure 4: Method of identifying the starting and ending points

decision. Therefore, we calculate the first and second derivatives for the individual subbands to obtain the speech period candidates.

We will use Fig. 4 to explain the mechanism for identifying the starting (Fig. 4a) and ending (Fig. 4b) points. These two figures show the observation frames. To determine the starting and ending points, the speech signal is observed in segments of 8 frames with an overlap of 4 frames. The rules for identifying the starting and ending points are as follows:

(i) To identify a starting point, we consider 8 frames at once and check the former 4 frames, as shown in Fig. 4a. We observe these frames to find a frame that has the local maximum second derivative followed by a positive first derivative in the successive frame. When this pattern holds, such frame becomes a starting point candidate. This process continues for the successive eight frames with an overlap of four frames from the previous observation.

(ii) To identify an ending point, we consider 8 frames at once and check the subsequent 4 frames, as shown in Fig. 4b. We observe these frames to find a frame that has the combination of a local minimum first derivative and a local maximum second derivative that is preceded by at least one negative first derivative. When this pattern holds, such frame becomes an ending point candidate. This process continues for the successive eight frames overlapped with four frames from the previous observation.

The above two processes continue until the last observation frames have been examined.

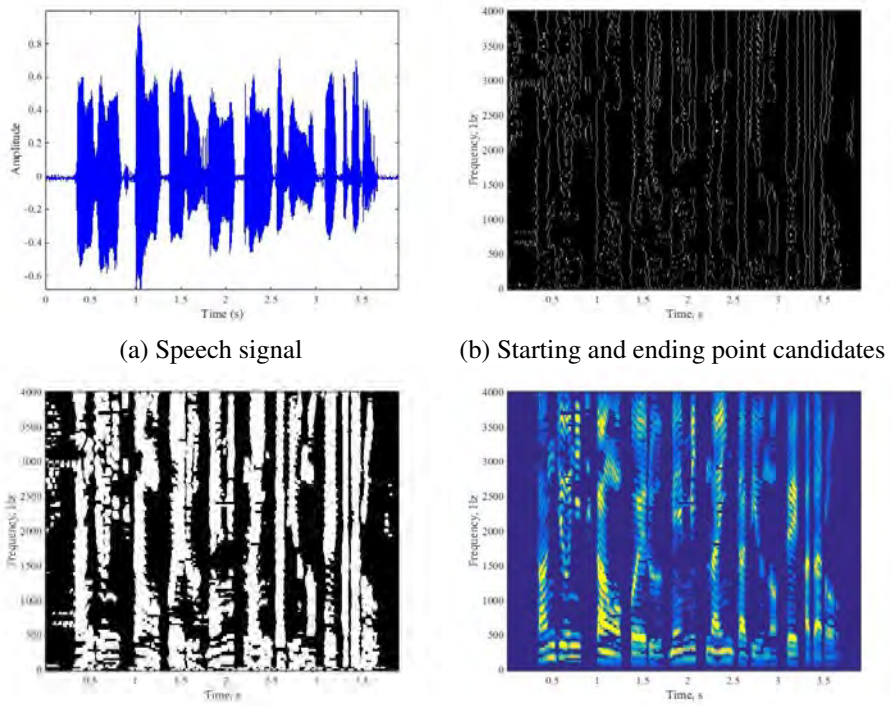
The starting and ending point candidates that are found based on rules (i) and (ii) are marked by the simple binary number of one. Figure 5b shows the starting and ending point candidates of the speech signal. We then simply add the binary ones between the starting and ending points to obtain the masks, as shown in Fig. 5c.

The masks, however, may cause misjudgments for non-speech periods because such masks do not carry information coming from the amplitude of the observed signal. To minimize such misjudgement, we attempt to remove values of 'one' coming from the signal parts when the amplitudes are relatively small simply by multiplying the power spectra expressed in decimal by the masks. Hereafter, the output of this process is referred to as speech period candidates. The result of the process is shown in Fig. 5d. These speech period candidates, together with the log power spectra from Eq. (2) become input for the DNN.

2.2 DNN based VAD

DNNs have been shown to be effective in various speech applications, including the detection of speech periods, as shown in [22]. According to [33], a DNN is a conventional multilayer perceptron (MLP) with many hidden layers. For simplicity, for an $L+1$-layer DNN, the input layer is regarded as layer 0, and the output layer is considered layer L. In the first L layers, an activation vector $\mathbf{a}^{\ell}$ is obtained as follows:

$$\mathbf{a}^{\ell} = f(\mathbf{z}^{\ell}) = f(\mathbf{W}^{\ell}\mathbf{a}^{\ell-1} + \mathbf{b}^{\ell}), \text{for} 0 < \ell < L. \tag{6}$$



(a) Speech signal (b) Starting and ending point candidates

(c) Masks (d) Speech period candidates

Figure 5: Representation of a speech signal (a), its starting and ending point candidates using rules (i) and (ii) (b), masks (c), and speech period candidates as a result of multiplying masks by the spectra expressed in decimal form (d)

where $\mathbf{z}^\ell \in \mathbb{R}^{\mathbf{N}_\ell \times 1}$ is an excitation vector, $\mathbf{W}^\ell \in \mathbb{R}^{\mathbf{N}_\ell \times \mathbf{N}_{\ell-1}}$ is a weight vector,$\mathbf{b}^\ell \in \mathbb{R}^{\mathbf{N}_\ell \times 1}$ is a bias vector, and $\mathbf{N}_\ell \in \mathbb{R}$ is the number of neurons in layer ℓ. $f(.) :\in \mathbb{R}^{N_\ell \times 1} \rightarrow \mathbb{R}^{N_\ell \times 1}$ is the activation function applied to the excitation vector element-wise $\mathbf{z}$. Here, the sigmoid function

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{7}$$

is used.

In the input layer, $\mathbf{a}^0 = \mathbf{P}(m,k) \in \mathbb{R}^{N_0 \times 1}$ denotes the input feature vector of the DNN, where $\mathbf{P}(m,k)$ denotes the log power spectra and the speech period candidates. $\mathbf{P}(m,k)$ is used as a learning instance and is mapped onto the correct speech periods that are identified during the training process.

In the output layer for classification tasks such as VAD, each output neuron represents a class $i \in \{1, \ldots, C\}$, where $C = N_L$ is the number of classes. In the output layer, a softmax function is added as a linear classifier for VAD:

$$\mathrm{a}_i^L = \mathrm{softmax}_i(\mathbf{z}^L) = \frac{e^{z_i^L}}{\sum_{j=1}^{C} e^{z_J^L}}, \tag{8}$$

where z_i^L is an ith element of the excitation vector $\mathbf{z}^L$. The value of the i-th output neuron a_i^L represents the probability $P_{dnn}(i|\mathbf{P})$ that the observation $\mathbf{P}(m,k)$ belongs to class i (speech or non-speech).

The training process for the DNN mentioned above consists of two stages. First, a greedy layer-wise unsupervised learning procedure is performed as the pre-training stage. Next, fine-tuning is performed on the entire network [34]. The DNN considered in this study is composed of five layers of restricted Boltzmann machines (RBMs), which consist of visible and hidden units. Here, Bernoulli (visible) − Bernoulli (hidden) RBMs, i.e., $v_\ell \in \{0,1\}$ and $h_\ell \in \{0,1\}$, are used. Once the learning process has been completed for an RBM, the activity values of its hidden units can be used as the *feature input* for training the next RBM [35]. A contrastive divergence algorithm is used in the pre-training stage to approximate the gradient of the negative log likelihood of the data with respect to the RBM's parameters [36]. After the layer-by-layer pre-training stage, a backpropagation technique is applied throughout the entire net to fine-tune the weights to obtain optimal results [37]. Because the VAD output consists of two classes (i.e. speech and non-speech), the DNN-based VAD output for each frame is a binary vector whose elements are determined as follows:

$$y_m = \mathbf{a}^L = \begin{cases} 1, & \text{if speech is at frame } m, \text{ and} \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

The DNN outputs trains of 1s (ones) representing the speech periods.

3 Experimental results and discussion

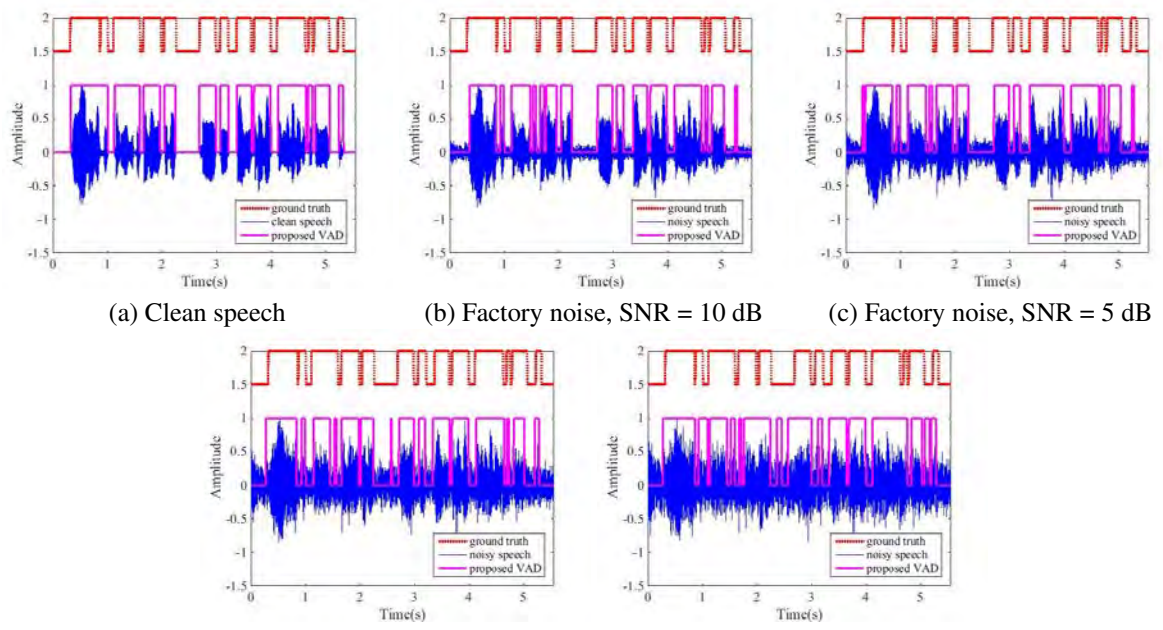
3.1 Experimental setup

In the experiments, we use 99 speech files from the ASJ Continuous Speech Corpus for Research vol. 2 [38]. These speech files are divided equally into 3 data sets. Then, we create three groups, each with 66 files for training (a combination of 2 data sets to obtain 3 distinct groups) and the rest of the speech files are used for evaluation purposes. The objective of dividing the data is to evaluate the proposed method inside a different data set. To obtain noisy signals, the clean speech files are mixed with 5 types of noise, white, babble, factory, car, and pink, from NOISEX-92 [39]. Each noise signal is differently selected for the speech files as well as SNRs of 10, 5, 0 and -5 dB. Thus, 21 sets of data are used to produce 1386 speech files for the training of each group.

In this work, the input signals are sampled at 8 kHz. The frame size is 20 ms, and the analysis window is a Hamming window with a 10 ms frame shift. After the RSF filtering process, the LPS-RSF is calculated for an individual subband using Eq. (3). Next, the first and second derivatives for each subband are calculated using Eqs. (4) and (5). These derivatives are used to obtain the starting and ending point candidates in accordance with rules (i) and (ii). After the conversion of the starting and ending point candidates from the sparse representation to the masks, the masks are multiplied by the power spectra, expressed in decimal form to obtain speech period candidates. Then, the DNN is applied to the speech period candidates in combination with the log power spectra derived from Eq. (2). This combination (i.e., the dynamic features) is fed to the DNN to obtain the final VAD decision regarding the speech periods. The dynamic features are normalized to zero mean and unit variance in each dimension. To train the DNN, we use five RBMs. These RBMs are stacked together, and the number of neurons for each RBM is 200, 200, 200, 200, and 100, in sequence. The learning rate is 0.0001, and the maximum number of epochs for both the pre-training and fine tuning stages is 200.

Note that, after determining the speech and non-speech periods, we do not perform any post processing, such as a VAD *hangover*, because such processing is outside the scope of this paper.

3.2 Results and discussion



(a) Clean speech (b) Factory noise, SNR = 10 dB (c) Factory noise, SNR = 5 dB

(d) Factory noise, SNR = 0 dB (e) Factory noise, SNR = -5 dB

Figure 6: Representative results of the proposed VAD method

The VAD decisions on noisy signals smeared with factory noise at various SNRs are shown in Fig. 6. In Fig. 6, the red dashed lines indicate the true speech and non-speech periods, whereas the solid magenta lines represent the generated VAD output. As shown in the figure, the output of the proposed VAD method is reasonably close to the ground truth.

To evaluate the effectiveness of the proposed method, we compare it with the DNN-based VAD method, which utilizes the log power spectra, as the baseline of the evaluation.

To represent the performance of the proposed method, the receiver operation characteristic (ROC) curve, in which true positive rate (TPR) is plotted against false positive rate (FPR), is considered. The TPR, or sensitivity, and FPR are defined based on the number of true positives (TPs), false positives (FPs), true negatives (TNs), and false negatives (FNs) as follows

[40]:

$$TPR = \frac{TP}{TP + FN}, \text{ and} \tag{10}$$

$$FPR = \frac{FP}{FP + TN}. \tag{11}$$

To obtain a quantitative ROC value, the area under the curves (AUCs) are calculated. These AUCs are the main metric for evaluation.

Table 1: AUC (%) comparison between the proposed method and DNN-based VAD methods using speech period candidates and log power spectra as the baseline. The numbers in bold indicate the best results

		AUC (%) - mean ± standard deviation		
Noise	SNR (dB)	Proposed	Log power spectra	Speech period candidates
Clean		**99.06 ± 0.13**	98.72 ± 0.20	98.10 ± 0.39
White	10	**97.91 ± 0.28**	97.51 ± 0.49	97.06 ± 0.54
	5	**97.44 ± 0.43**	97.27 ± 0.48	96.64 ± 0.46
	0	**96.59 ± 0.50**	96.14 ± 0.76	95.44 ± 0.57
	-5	**94.69 ± 0.66**	93.88 ± 1.10	93.40 ± 0.60
Babble	10	**96.84 ± 0.60**	96.50 ± 0.55	96.19 ± 0.68
	5	**95.19 ± 0.71**	94.26 ± 0.66	94.59 ± 0.92
	0	**91.30 ± 0.74**	88.88 ± 0.74	90.42 ± 0.53
	-5	**83.20 ± 0.87**	78.10 ± 1.10	81.85 ± 0.85
Factory	10	**97.25 ± 0.39**	96.80 ± 0.60	96.60 ± 0.56
	5	**95.96 ± 0.43**	95.14 ± 0.72	95.48 ± 0.77
	0	**93.18 ± 0.46**	91.17 ± 0.45	92.53 ± 0.67
	-5	**85.91 ± 0.29**	80.49 ± 1.54	84.57 ± 0.83
Car	10	**99.02 ± 0.11**	98.83 ± 0.15	97.60 ± 0.45
	5	**98.94 ± 0.11**	98.75 ± 0.16	97.37 ± 0.45
	0	**98.79 ± 0.09**	98.56 ± 0.16	97.02 ± 0.41
	-5	**98.40 ± 0.05**	98.06 ± 0.02	96.36 ± 0.32
Pink	10	**97.79 ± 0.39**	97.20 ± 0.66	96.86 ± 0.73
	5	**96.82 ± 0.59**	96.28 ± 0.79	95.98 ± 0.74
	0	**95.26 ± 0.70**	94.06 ± 0.95	94.26 ± 0.89
	-5	**91.56 ± 1.03**	88.01 ± 1.54	89.91 ± 1.20

Table 1 compares the average of the AUCs achieved by the DNN-based VAD methods using the log power spectra alone, the speech period candidates alone, and the proposed method. As shown in Table 1, the proposed method may improve the performance of the DNN-based VAD method using the log power spectra for all cases. As shown in Fig. 7, a relatively good improvement is obtained in low SNR cases. The highest improvement occurs at -5 dB, for example, the performance improves by 6.52% for babble noise.

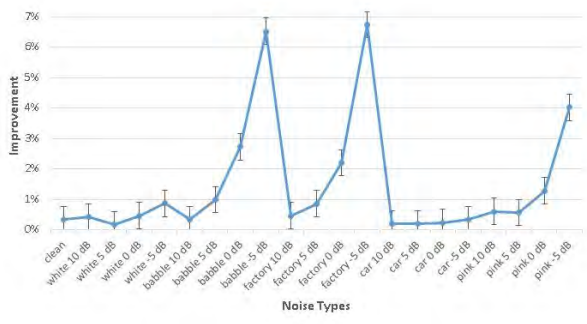
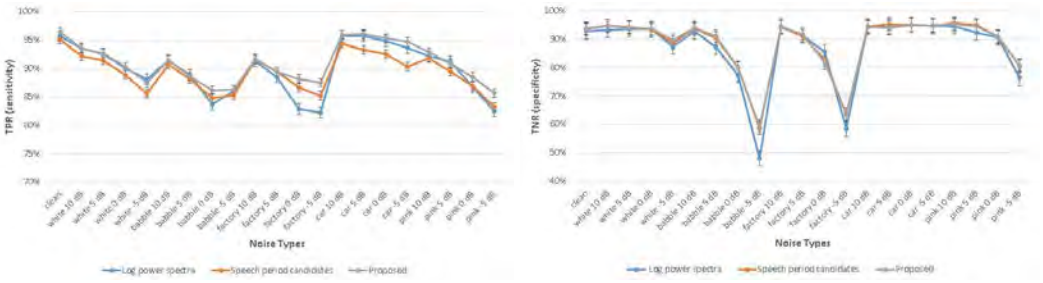


Figure 7: Improvement of VAD performance

To evaluate the effect of introducing the speech period candidates, we measure the TPR (sensitivity) and true negative rate (TNR or specificity). Sensitivity gives the percentage of the frames that were correctly classified as speech from all the

speech frames in the signal and specificity gives the percentage of the frames that were correctly classified as non-speech from all the non-speech frames in the signal [41].

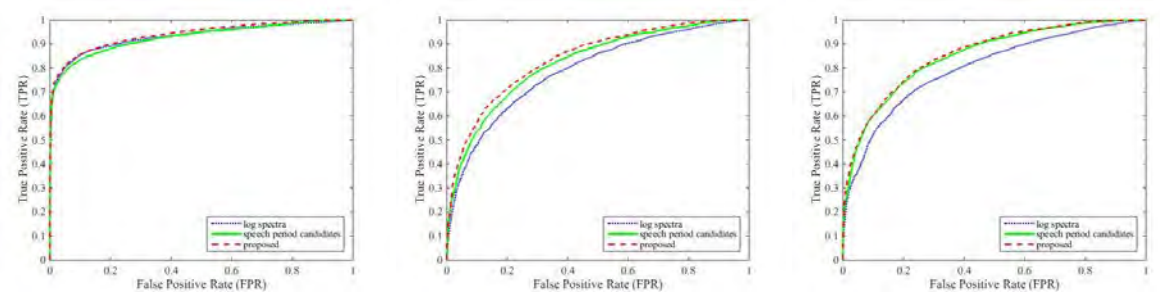
Figure 8 shows the mean TPR (sensitivity) and TNR (specificity). As shown in the figure, the proposed method has a high sensitivity and specificity for both the clean and noisy cases. Interestingly, the performance of the DNN-based VAD method using speech period candidates approaches and even outperforms the log power spectra in finding speech, as shown in Fig. 8a, particularly for low SNRs and non-stationary cases. This fact may imply that the addition of speech period candidates is useful to find speech periods in low SNRs and non-stationary cases. The specificity of speech period candidates is higher than the log power spectra as shown in Fig. 8b. This fact may imply that the speech period candidates may improve the log power spectra for finding non-speech periods. Thus, the speech period candidates may carry valuable information for judging speech and non-speech detection. In the proposed method, the addition of the speech period candidates is effective at improving the accuracy of the log power spectra at finding speech and non-speech periods, especially for low SNRs and non-stationary cases.



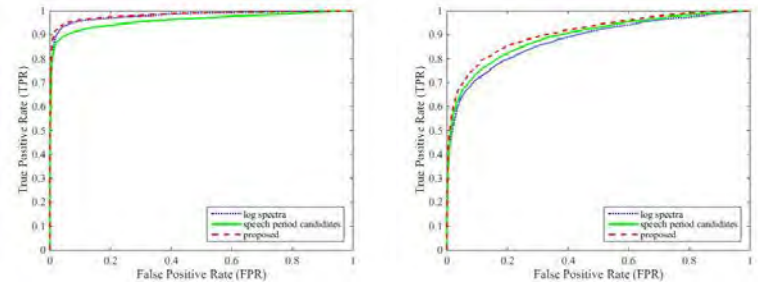
(a) TPR (sensitivity) (b) TNR (specificity)

Figure 8: Sensitivity and specificity comparison between the proposed method, and the DNN-based VAD methods using speech period candidates and log power spectra

Figure 9 shows the ROC curves for the proposed method and the DNN-based VAD methods using speech period candidates and log power spectra, respectively, at an SNR of -5 dB. As shown in the figure, the proposed method shows an advantage over the DNN-based VAD methods. The proposed method is effective for low SNR cases. In the cases of stationary noise, such as white and pink noise, the working points of the proposed method are close to those of the DNN-based VAD method using the log power spectra. These methods achieve a high TPR and a low FPR. In the cases of non-stationary noise, such as babble and factory noise, the proposed method is less affected by the noise than the DNN-based VAD method using the log power spectra. The performance of the proposed method is superior to that of the DNN-based VAD method using the log power spectra mainly due to introducing dynamics expressed by speech period candidates.



(a) White noise, SNR = -5 dB (b) Babble noise, SNR = -5 dB (c) Factory noise, SNR = -5 dB



(d) Car noise, SNR = -5 dB (e) Pink noise, SNR = -5 dB

Figure 9: ROC curves for the proposed method and DNN-based VAD methods using log power spectra and speech period candidates, respectively

In addition to the contribution of the speech period candidates, which may highlight dynamics, we attempt to find useful subbands for obtaining VAD decisions in the employed DNN. We evaluate which subbands have more valuable information than the others, by finding the similarity between the input (i.e., speech period candidates) and the VAD output. This similarity is evaluated by employing mutual information (MI), which aims to measure whether the inputs are dependent on the associated labels (VAD output). According to [42], the MI between the discretized feature values, a, and the class labels, y, is evaluated according to the formula

$$MI = \sum_{a \in A} \sum_{y \in Y} p(a, y) \log(\frac{p(a, y)}{p(a)p(y)}), \tag{12}$$

where $p(a, y)$ is a joint probability function of a and y, and $p(a)$ and $p(y)$ are marginal probability distribution functions of a and y, respectively. Here, the features values, a, are the input of the DNN (speech period candidates), and the class labels, y, are the VAD output. The larger the MI, the higher the dependency between the feature values, which represent speech period candidates for individual subbands, and the class labels (VAD output). Here, we rank the subbands according to their scores.

Table 2: Subband (Hz) ranks using mutual information (MI)

		Subband (Hz) ranks using MI			
Noise	SNR (dB)	1	2	3	4
Clean		187.5	218.75	156.25	312.5
White	10	187.5	218.75	156.25	250
	5	187.5	218.75	156.25	250
	0	187.5	218.75	156.25	250
	-5	187.5	218.75	156.25	250
Babble	10	187.5	218.75	250	156.25
	5	187.5	218.75	250	156.25
	0	187.5	218.75	250	156.25
	-5	187.5	218.75	250	156.25
Factory	10	187.5	218.75	156.25	250
	5	187.5	218.75	156.25	250
	0	187.5	218.75	250	156.25
	-5	218.75	187.5	250	312.5
Car	10	187.5	218.75	156.25	250
	5	187.5	218.75	312.5	250
	0	187.5	218.75	250	312.5
	-5	218.75	187.5	250	312.5
Pink	10	187.5	218.75	156.25	250
	5	187.5	218.75	250	156.25
	0	187.5	218.75	250	156.25
	-5	187.5	218.75	250	156.25

Table 2 shows the top 4 subband ranks using MI. As shown in Table 2, the top 4 ranks for clean and noisy signals show a similar tendency for frequency bins 6, 7, 8, and 9 (156.25 Hz, 187.5 Hz, 218.75 Hz, and 250 Hz). Such subband may play some roles in obtaining the VAD decision in the proposed method. To clarify this, we perform experiments in which the 4 top subband values in the proposed method are replaced with zeros, and the resulting VAD performance is shown in Fig. 10.

As shown in Fig. 10, at a high SNR, the performance of the proposed method is only slightly degraded when the subbands of 156.25 Hz, 187.5 Hz, 218.75 Hz, and 250 Hz are replaced by zeros. In low SNR cases, the subbands are polluted by noise. Consequently, the performance might be degraded, and this degradation worsens when these top 4 subbands are not utilized. In contrast, when the 4 lowest subband values are replaced by zeros, the output accuracy can still be maintained. These results indicate that the top 4 subbands have a relatively important role in the decision-making process of the proposed method. We observe that the information carried by these subbands may correspond to the average of F0 or its neighbors (average F0 for the data is 149.09 Hz for male and 210.84 Hz for female). Thus, in the proposed method, the DNN may utilize information coming from the useful subbands which may correspond to F0 or its neighbors.

4 Conclusions

This study presents a DNN-based VAD method for improving the performance of VAD by introducing dynamics which may be highlighted by speech period candidates. These candidates are derived from heuristic rules based on the first and second derivatives of the log power spectrum of the RSF output (LPS-RSF). The speech period candidates are calculated

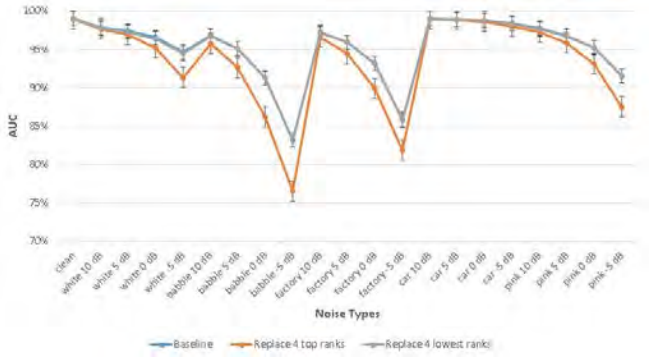


Figure 10: VAD performance of the proposed method after replacing the subband values of top 4 ranks' and the lowest 4 ranks' with zeros

for individual subbands. These candidates are then input into a DNN together with the log power spectra to generate the VAD decision. To evaluate the performance of the proposed method, we perform experiments using clean and noisy speech signals smeared with five types of noise, namely, white, babble, factory, car and pink, with SNRs of 10, 5, 0 and -5 dB. The proposed method effectively detects speech and non-speech periods. The experimental results show that the VAD performances based on log power spectra are improved after combining the log power spectra with the speech period candidates, particularly for noisy speech signals with low SNRs and non-stationary cases. The addition of dynamics expressed by the speech period candidates provides positive information that contributes to the detection of speech periods.

In this study, we also show that the DNN-based VAD utilizes subbands that may correspond to F0 or its neighbours. These subbands may also be less affected by the noise. The VAD performance degrades when those subbands are eliminated. However, further studies should be performed to analyze other factors that influence the behaviour of the employed DNN. Moreover, we intend to make the proposed method works in real time.

References

[1] K. Sakhnov, E. Verteletskaya, and B. Simak, "Approach for energy-based voice detector with adaptive scaling factor." *IAENG International Journal of Computer Science*, vol. 36, no. 4, 2009.

[2] F. Beritelli, S. Casale, and G. Ruggeri, "Performance evaluation and comparison of itu-t/etsi voice activity detectors," in *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, vol. 3. IEEE, 2001, pp. 1425–1428.

[3] A. Benyassine, E. Shlomot, H. Su, D. Massaloux, C. Lamblin, and J. Petit, "Itu-t recommendation g. 729 annex b: a silence compression scheme for use with g. 729 optimized for v. 70 digital simultaneous voice and data applications," *IEEE Communications Magazine*, vol. 35, no. 9, pp. 64–73, 1997.

[4] S. Tong, N. Chen, Y. Qian, and K. Yu, "Evaluating vad for automatic speech recognition," in *Signal Processing (ICSP), 2014 12th International Conference on*. IEEE, 2014, pp. 2308–2314.

[5] S. Graf, T. Herbig, M. Buck, and G. Schmidt, "Features for voice activity detection: a comparative analysis," *EURASIP Journal on Advances in Signal Processing*, vol. 2015, no. 1, p. 91, 2015.

[6] R. Rabiner and R. Sambur, "An algorithm for determining the endpoints of isolated utterances," *Bell Labs Technical Journal*, vol. 54, no. 2, pp. 297–315, 1975.

[7] R. Prasad, A. Sangwan, H. Jamadagni, M. Chiranth, R. Sah, and V. Gaurav, "Comparison of voice activity detection algorithms for voip," in *Computers and Communications, 2002. Proceedings. ISCC 2002. Seventh International Symposium on*. IEEE, 2002, pp. 530–535.

[8] J. Ramırez, J. C. Segura, C. Benıtez, A. D. L. Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech communication*, vol. 42, no. 3, pp. 271–287, 2004.

[9] K. Ishizuka, T. Nakatani, M. Fujimoto, and N. Miyazaki, "Noise robust voice activity detection based on periodic to aperiodic component ratio," *Speech Communication*, vol. 52, no. 1, pp. 41–60, 2010.

[10] K. Pek, T. Arai, and N. Kanedera, "Voice activity detection in noise using modulation spectrum of speech: Investigation of speech frequency and modulation frequency ranges," *Acoustical Science and Technology*, vol. 33, no. 1, pp. 33–44, 2012.

[11] T. Kinnunen and P. Rajan, "A practical, self-adaptive voice activity detector for speaker verification with noisy telephone and microphone data." in *ICASSP*, 2013, pp. 7229–7233.

[12] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE signal processing letters*, vol. 6, no. 1, pp. 1–3, 1999.

[13] A. Davis, S. Nordholm, and R. Togneri, "Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 2, pp. 412–424, 2006.

[14] T. Kinnunen, E. Chernenko, M. Tuononen, P. Fränti, and H. Li, "Voice activity detection using mfcc features and support vector machine," in *Int. Conf. on Speech and Computer (SPECOM07), Moscow, Russia*, 2007.
[15] Q. Jo, J. H. Chang, J. Shin, and N. Kim, "Statistical model-based voice activity detection using support vector machine," *IET Signal Processing*, vol. 3, no. 3, pp. 205–210, 2009.
[16] D. Enqing, L. Guizhong, Z. Yatong, and Z. Xiaodi, "Applying support vector machines to voice activity detection," in *Signal Processing, 2002 6th International Conference on*, vol. 2. IEEE, 2002, pp. 1124–1127.
[17] G. Ferroni, R. Bonfigli, E. Principi, S. Squartini, and F. Piazza, "A deep neural network approach for voice activity detection in multi-room domestic scenarios," in *Neural Networks (IJCNN), 2015 International Joint Conference on*. IEEE, 2015, pp. 1–8.
[18] X. Zhang and D. Wang, "Boosting contextual information for deep neural network based voice activity detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2016.
[19] F. Bie, Z. Zhang, D. Wang, and T. Zheng, "Dnn-based voice activity detection for speaker recognition," Tech. Rep, Tech. Rep., 2015.
[20] A. Mohamed, G. Hinton, and G. Penn, "Understanding how deep belief networks perform acoustic modelling," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 4273–4276.
[21] M. Espi, M. Fujimoto, K. Kinoshita, and T. Nakatani, "Exploiting spectro-temporal locality in deep learning based acoustic event detection," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, p. 26, 2015.
[22] N. Ryant, M. Liberman, and J. Yuan, "Speech activity detection on youtube using deep neural networks." in *INTERSPEECH*, 2013, pp. 728–731.
[23] V. Mendelev, T. Prisyach, and A. Prudnikov, "Robust voice activity detection with deep maxout neural networks," *Modern Applied Science*, vol. 9, no. 8, p. 153, 2015.
[24] X. Zhang and J. Wu, "Deep belief networks based voice activity detection," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 4, pp. 697–710, 2013.
[25] L. Deng, J. Li, J. Huang, K. Yao, D. Yu, F. Seide, M. Seltzer, G. Zweig, X. He, J. Williams, Y. Gong, and A. Acero, "Recent advances in deep learning for speech research at microsoft," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8604–8608.
[26] L. Deng, "Dynamic speech models: theory, algorithms, and applications," *Synthesis Lectures on Speech and Audio Processing*, vol. 2, no. 1, pp. 1–118, 2006.
[27] K. Fujioka, N. Hayasaka, Y. Miyanaga, and N. Yoshida, "Noise reduction of speech signals by running spectrum filtering," *Systems and Computers in Japan*, vol. 37, no. 14, pp. 52–61, 2006.
[28] N. Kanedera, T. Arai, H. Hermansky, and M. Pavel, "On the relative importance of various components of the modulation spectrum for automatic speech recognition," *Speech Communication*, vol. 28, no. 1, pp. 43–55, 1999.
[29] H. Hermansky and N. Morgan, "Rasta processing of speech," *IEEE transactions on speech and audio processing*, vol. 2, no. 4, pp. 578–589, 1994.
[30] L. Atlas and S. Shamma, "Joint acoustic and modulation frequency," *EURASIP Journal on Advances in Signal Processing*, vol. 2003, no. 7, p. 310290, 2003.
[31] T. Arai, M. Pavel, H. Hermansky, and C. Avendano, "Syllable intelligibility for temporally filtered lpc cepstral trajectories," *The Journal of the Acoustical Society of America*, vol. 105, no. 5, pp. 2783–2791, 1999.
[32] Q. Zhu, N. Ohtsuki, Y. Miyanaga, and N. Yoshida, "Robust speech analysis in noisy environment using running spectrum filtering," in *Communications and Information Technology, 2004. ISCIT 2004. IEEE International Symposium on*, vol. 2. IEEE, 2004, pp. 995–1000.
[33] D. Yu and L. Deng, *Automatic speech recognition: A deep learning approach*. London: Springer, 2014.
[34] H. Larochelle, Y. Bengio, J. Louradour, and P. Lamblin, "Exploring strategies for training deep neural networks," *Journal of Machine Learning Research*, vol. 10, no. Jan, pp. 1–40, 2009.
[35] G. Hinton, "Learning multiple layers of representation," *Trends in cognitive sciences*, vol. 11, no. 10, pp. 428–434, 2007.
[36] M. Carreira-Perpinan and G. Hinton, "On contrastive divergence learning." in *Aistats*, vol. 10, 2005, pp. 33–40.
[37] M. Keyvanrad and M. Homayounpour, "A brief survey on deep belief networks and introducing a new object oriented toolbox (deebnet)," *arXiv preprint arXiv:1408.3264*, 2014.
[38] T. Kobayashi, "ASJ continuous speech corpus for research," *Acoustic Society of Japan (ASJ) Trans*, vol. 48, no. 12, pp. 888–893, 1992.
[39] Noisex-92 database. [Online]. Available: http://spib.linse.ufsc.br/noise.html
[40] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. John Wiley & Sons, 2012.
[41] M. Myllymäki and T. Virtanen, "Voice activity detection in the presence of breathing noise using neural network and hidden markov model," in *Signal Processing Conference, 2008 16th European*, 2008, pp. 1–5.
[42] J. Pohjalainen, O. Räsänen, and S. Kadioglu, "Feature selection methods and their combinations in high-dimensional classification of speaker likability, intelligibility and personality traits," *Computer Speech & Language*, vol. 29, no. 1, pp. 145–171, 2015.

平成 30 年　2 月　6 日

学位論文審査報告書（甲）

1．学位論文題目（外国語の場合は和訳を付けること。）

Voice Activity Detection Using Deep Neural Network

（ディープニューラルネットワークを用いた音声区間検出）

2．論文提出者　(1) 所　　属　電子情報化学専攻　専攻

(2) 氏（ふり）　名（がな）　Dwijayanti Suci（ドウィジャヤンティ　スチ）

3．審査結果の要旨（600～650 字）

平成 30 年 2 月 6 日 10:00-10:30 に第 1 回学位論文審査会を開催し，同日 10:30-11:30 に口頭発表会を実施した．その直後に第 2 回審査委員会を開き，慎重に審議し，以下の通り判定した．なお，口頭発表における質疑を最終試験に代えた．

音声区間検出（VAD）は種々の音声アプリケーションの前処理として，雑音重畳した入力信号中の音声区間と非音声区間を検出するために用いられる．これまで，検出に有用な音声特徴量とその利用方法が様々検討されて来た．本研究では，まず，入力信号の対数パワスペクトル（LPS）と制限ボルツマンマシン（RBM）を用いるディープニューラルネットワーク（DNN）を組み合わせた VAD の性能が，従来の代表的な方法を上回ることを実験的に示した．次に，従来技術が，音声特徴量の時間変化を表す量を加えて当該特徴量を単独で用いる場合の VAD 性能を改善していることを考慮し，LPS の時間変化を強調する補助特徴量（Speech Period Candidates）を考案し，LPS と組み合わせる方法を開発した．実験の結果，提案方法は，特に，SN 比の悪い（-5～0 dB）非定常雑音が重畳された場合に，LPS を単独で用いる場合の VAD 性能を大幅に改善出来ることが示された．また，DNN は音声の基本周波数と近傍を VAD の重要な手掛かりにしていることも確認した．以上より，本論文は博士（工学）に値すると判定した．

4．審査結果　(1) 判　　定（いずれかに○印）　○合　格　・　不合格

(2) 授与学位　　博　士（　工学　）